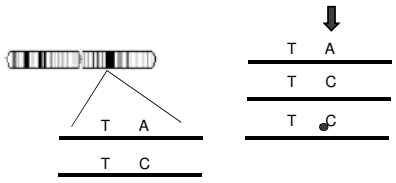
Complex Disorders: Populations and Pedigrees (II)

CGP Lecture, 20th October 2015
Dr CA Sargent
cas1001@cam.ac.uk

Principles of Association analysis



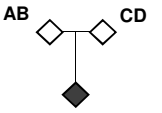
- Simple case: two haplotypes, TA and TC at two adjacent loci
- Polymorphism contributing to trait lies near C on TC background
- Separate into cases and controls.
- Genotype at locus A/C

Spurious association

- **Disease susceptibility shows ethnic differences**
- **Polymorphisms show ethnic differences**
- **This can lead to spurious or false associations being detected if samples from different populations are analysed together**

Transmission Disequilibrium Test (TDT)

Case-control study: is allele A more frequent in cases than controls?



Cases + parents

TDT: when a parent has allele A and is heterozygous, is allele A transmitted to the affected offspring more frequently than the expected 50% of times?

TDT avoids spurious association due to population stratification

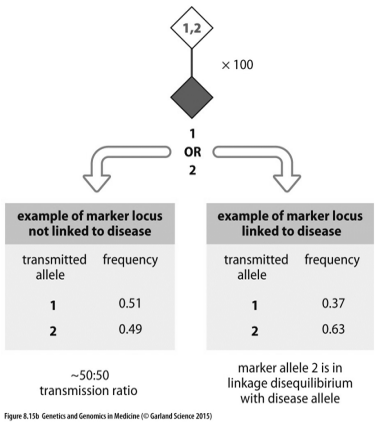


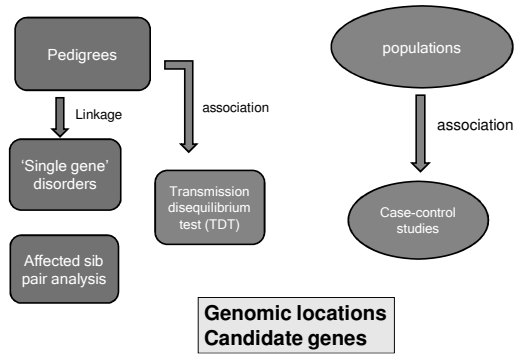
Figure 8.15b Genetics and Genomics in Medicine (© Garland Science 2015)

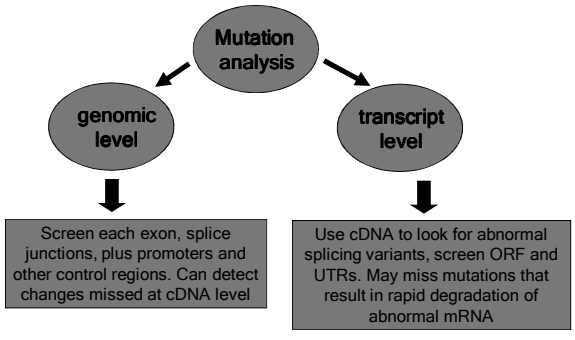
Linkage vs association mapping

- **Linkage detectable over large genetic distances, typically 10-20 cM with large sample and many informative meioses**
- **Allelic association has to persist over many generations, so only detectable over small genetic distances of the order of 1 cM (humans)**
- **Linkage and association complement each other**
- **New methods of combined linkage and LD analysis enable the two approaches to be combined in a single analysis**
  - **Especially useful with multiple small pedigrees with >1 affected sib**

Linkage and association pros and cons

- **Linkage studies**
  - **Good power for genes of large to medium effect**
  - **Need extremely large samples to detect weak effects**
  - **Computationally difficult with large number of genotypes**
- **Association studies**
  - **More powerful than linkage for small gene effects**
  - **Suitable for high throughput genotyping**
  - **Need to beware of spurious associations**

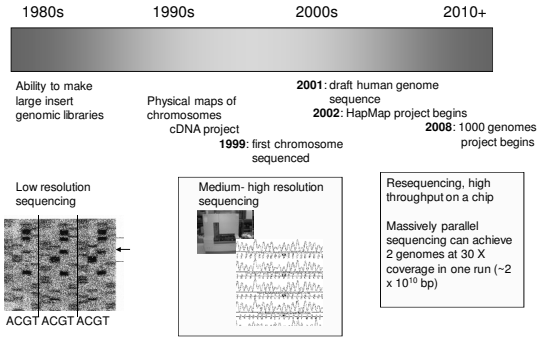




Polymorphic variant or disease causing variant?

- Genome sequences vary between individuals
- Which variants are truly 'polymorphic', with no contribution to phenotype, and which are not?

Sequencing progress aids SNP discovery



Genome sequence and structure

- 'In the initial phase of the Project, genetic data are being gathered from four populations with African, Asian, and European ancestry.'
- 'The goal of the International HapMap Project is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared'
- *Aims- define all polymorphisms where minor alleles are normally >5%, define blocks of linkage disequilibrium in the populations*

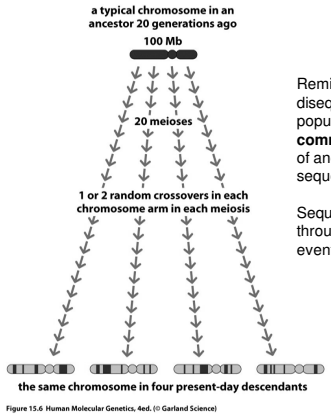
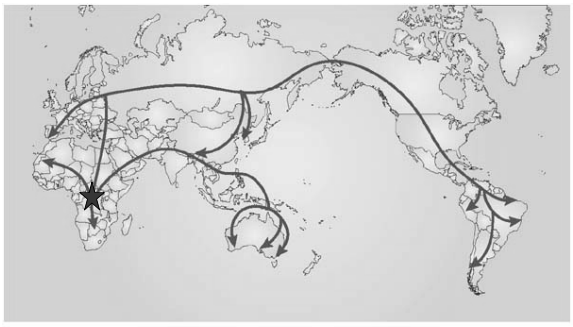


Figure 15.6 Human Molecular Genetics, 4ed, (© Garland Science)

Reminder: linkage disequilibrium blocks at the population level = the **common** shared segments of ancestral chromosomal sequences.

Sequences are reassorted through recombination events over generations.

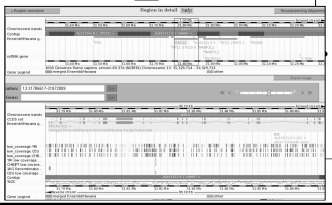
LD blocks mirror human migrations: they are shortest in the oldest populations (African) and longest in the newest (e.g. admixed populations). This reflects the number of generations from the first individuals to the extant individuals



Migration routes of *Homo sapiens*

Nature Reviews | Cancer

1000 genomes project



- 1. 2500 individuals collected
  - Look for rarer variants
  - Population differences
- 2. Parent-child trios
  - Look at de novo mutation rate
- 3. 1000 gene regions in depth
  - Accumulate more information about disease related regions of genome

BRCA2 region

First phase results (2010)


ARTICLE

A map of human genome variation from population-scale sequencing

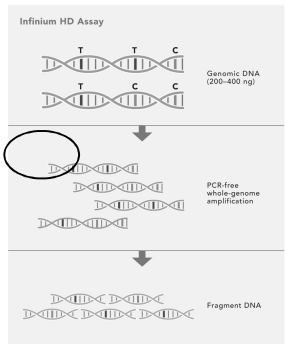
The 1000 Genomes Project Consortium*

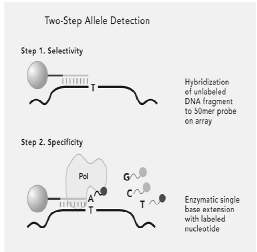
- **15 million** SNPs
  - 1/ 200bp
- **1 million** short indels
  - 1/ 3Kbp
- **20,000** structural variants
- Can detect 95% of differences between any two individuals
- We each carry
  - **250-300** loss of function mutations: so most genes do not show dosage effects (no haploinsufficiency)
  - **50-100** variants in genes previously implicated in genetic diseases
  - **True germline mutation rate is 10^{-8} per bp per generation** (about 60 new DNA mutations per generation)- fewer than previous estimates.

Using SNPs

- Easy to type
- Know where they are
- Know frequency and distribution of genotypes in different populations
  - SNPs need to show Hardy Weinberg equilibrium in the population as a whole, but most useful SNPs will have a minor allele frequency >5%
    - (*SNPs that don't show HW may already be under some form of selection. As SNPs approach MAF 50%, deviations from 50% provide greater power to the statistical analysis*)
- Experiments can be adapted to refine the experimental design
  - First pass experiments may use evenly spaced SNPs to get a genome-wide assessment, and further experiments use high density markers from implicated regions of the genome to confirm the initial observations.

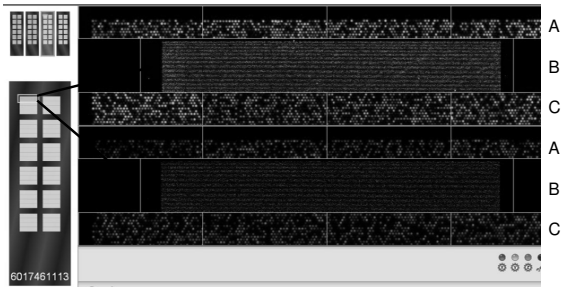
Principles of scoring SNPs: companies vary in technology

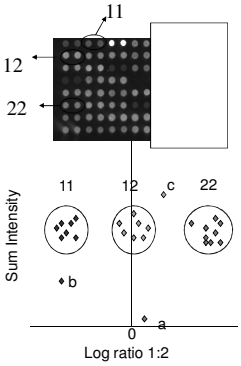




Use ddNTPs (chain terminating extension)
As C>T and G>A are commonest changes, focus on these

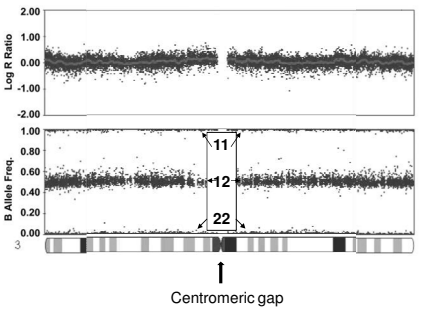
- The slide is scanned in two channels: one for each of the fluorescent dyes corresponding to specific bases at the SNP position
- Section B shows the strip across the slide
- Sections A and C are magnified images showing the beads.





- Once the two images are superimposed, the genotype can be read by the scanner
- The total intensity per bead should be within defined limits
- The ratio of the intensities in each channel confirm whether the marker is homozygous or heterozygous
  - Log ratios help to separate out the results for easier visualisation
- Samples that fall outside the appropriate ranges for ratio or intensity are either failed (a), or can be plotted to look for copy number variation (b,c)

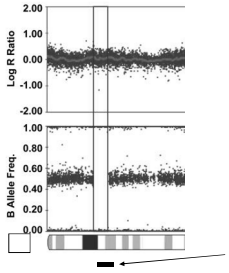
Results from across a chromosome: output for one individual using a SNP array

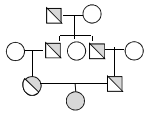


Log_2R ratio across chromosome in the upper panel is based on total fluorescence per position (0= normal copy number of 2 alleles)

The lower panel calls the genotype at each position. Most individuals will have an average representing heterozygosity across the length of the chromosome

Extended homozygosity in consanguinous marriages: autozygosity mapping





•In this study, copy number is correct, ie two per locus.
•However, long stretch of homozygosity.
•Patient is offspring of cousin-cousin marriage, and shows **autozygosity** at this position of genome.
•*selection of candidate genes by studying gene content of region*
•*Has been a good approach for discovery of obesity genes*

Complex disease: Hypothesis

- **There is an assumption that common disease arises owing to inheritance of common alleles that cumulatively have an impact on phenotype**
  - **(is this true?)**
- Case-control studies are a potential way forward with polygenic disorders, and are now carried out on a genome wide level, rather than by individual candidate genes.
- Because all association relies on LD, statistically significant alleles in a genome wide analysis should lie close to causative changes in the DNA sequence
- Considerations:
  - Reflection of population history e.g. admixture, bottlenecks, generation time, population size
  - is LD the same in all populations?
  - Need at least 3-5 times as many SNPs as MS for a genome wide study
    - *All above now feasible as result of genome studies*

The Wellcome Trust Case Control Consortium Study

- Used case control study with whole genome association based on SNP chips to study diseases with major economic impact
  - **Diabetes (types I and 2)**
  - **Bipolar disorder**
  - **Crohn's disease**
  - **Coronary artery disease**
  - ***Hypertension***
  - **Rheumatoid arthritis**

These are mainly discontinuous, or threshold, traits

(*Nature 2007: volume 447, p661*)
(*Genomics. 2008: 92, p265-72*)

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

2007: first major study
2000 cases per disease/ 3000 controls

ARTICLES

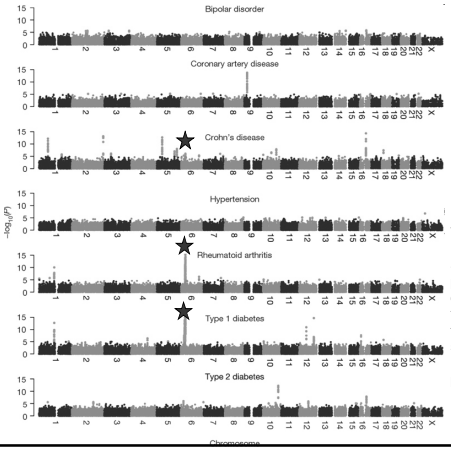
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

ARTICLES

nature genetics

2010: update on T2D
>40,000 cases/ 100,000 controls
Increase chance of detecting loci with small effects

Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis



Very few genes appear to have a major genetic contribution

Signal in MHC region for autoimmune diseases ★

The plots are $-\log_{10}p$ values against genomic position. Once a correction is made for the number of tests, genome wide significance ($p<0.05$) becomes $p<10^{-8}$ for 1 million markers.

Can we improve our data?

- Look at the bigger picture: can results be grouped into genetic pathways?
- Recognise that complex diseases show genetic heterogeneity, and work with multiple pedigrees
- Turn to quantitative aspects of discontinuous traits
  - These are the **endophenotypes**
  - E.g diabetes: BMI, serum leptin levels, serum insulin levels, triglyceride levels etc....

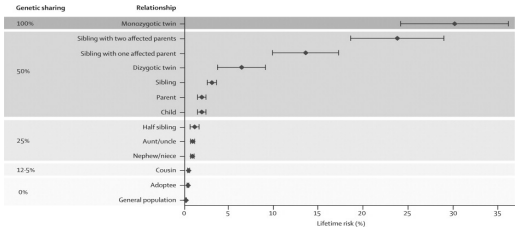
Case study I: Multiple Sclerosis - using data to define pathways and the nature of pathogenesis

1. Shows world-wide distribution differences (genetic or environmental?)

Geography of multiple sclerosis and migrations The five continents are depicted to show medium prevalence of multiple sclerosis (orange), areas of exceptionally high frequency (red), and those with low rates (grey-blue). Following migration reveals that a move from a low risk region to one of high risk increases risk over the parental population (solid arrows). Conversely, migration from high risk to lower risk areas decreases the risk over the parental population (dotted arrows). Implicates environmental factors.

Alastair Compston, Alasdair Coles
Multiple sclerosis The Lancet Volume 372, Issue 9648 2008 1502 – 1517 http://dx.doi.org/10.1016/S0140-6736(08)61620-7

2. Shows heritability (genetic effect)

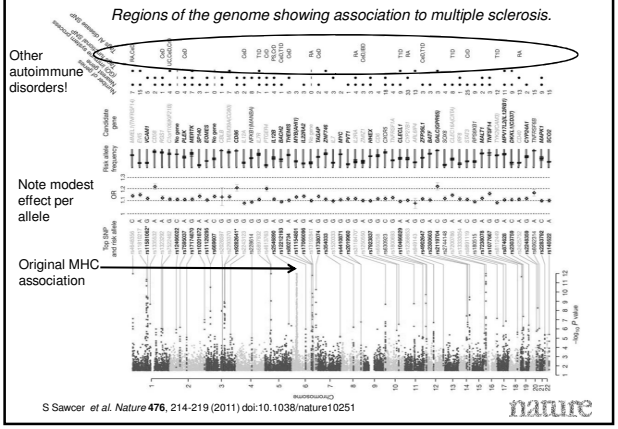


Recurrence risks for multiple sclerosis in families Age-adjusted recurrence risks for different relatives of probands with multiple sclerosis, and degree of genetic sharing between relative and proband. The graph indicates that risk increases with increased sharing of genetic material: i.e. there is a genetic component to the disease.

Alastair Compston, Alasdair Coles
Multiple sclerosis The Lancet Volume 372, Issue 9648 2008 1502 – 1517 http://dx.doi.org/10.1016/S0140-6736(08)61620-7

3. Is it neurodegenerative or inflammatory disorder?

Regions of the genome showing association to multiple sclerosis.

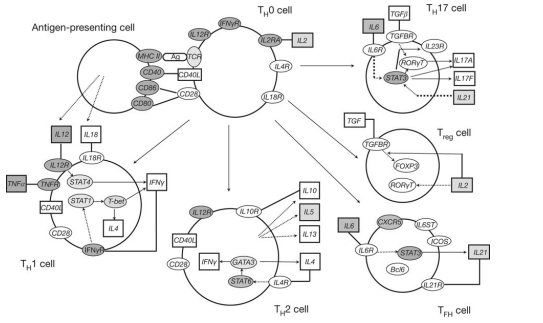


S Sawcer *et al. Nature* **476**, 214-219 (2011) doi:10.1038/nature10251

nature

4. Can consolidate data based on functional pathways

Graphic representation of the T-helper-cell differentiation pathway. Coloured nodes represent genes near significant SNPs that lie in pathways implicated in the disease.



S Sawcer *et al. Nature* **476**, 214-219 (2011) doi:10.1038/nature10251

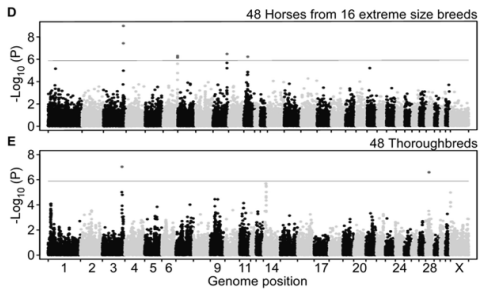
nature

Can we improve our data?

- Look at the bigger picture: can results be grouped into genetic pathways?
- Recognise that complex diseases show genetic heterogeneity, and work with multiple pedigrees
- Turn to quantitative aspects of discontinuous traits
  - These are the **endophenotypes**
  - E.g diabetes: BMI, serum leptin levels, serum insulin levels, triglyceride levels etc....

Height is a quantitative trait.....

Figure 1. Two genome-wide association scans for size identify five significantly associated loci.

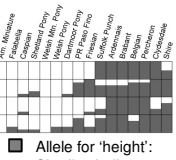


Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, et al. (2012) Four Loci Explain 83% of Size Variation in the Horse. PLoS ONE 7(7): e39929. doi:10.1371/journal.pone.0039929
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0039929

PLOS | ONE

Chr.	Position	Alleles Little	Alleles Big	P-value	Fraction of Variance Explained	Genes in Locus
3	105,547,002	T	C	1.05×10^{-8}	0.685	*LCORL NCAPG*
6	81,481,064	C	T	5.21×10^{-7}	0.556	*HMGA2*
9	75,550,059	C	T	3.48×10^{-7}	0.567	*ZFAT*
11	23,259,732	G	A	6.19×10^{-7}	0.589	*LASP1*
4 Loci Together					0.835	

From: PLoS One. 2012; 7(7): e39929.



Allele for 'height': Shading indicates allele frequency in the breed

Only four loci explain almost all of the variation in size in horse breeds (~83%). These have all been implicated in human height. *LCORL* is also involved in cattle growth, and *HMGA2* has been identified as associated with size in dog breeds.
In humans >200 height/size loci have been identified to date!

Combining GWAS with linkage and affected sib pair studies – quantitative phenotypes in diabetes

Table 3 • Markers with evidence for linkage

Marker	Chr	Location (cM from p-ter)	IBD	MLS
D10S197	10	55	57%	4.85
DXS1226	X	37	-	2.42
D6S344	6	1	55%	2.40
D2S206	2	273	53%	1.60
DXS1001	X	149	-	1.42
D7S517	7	8	55%	1.27
D5S647	5	81	54%	1.57
D5S436	5	159	53%	1.19
D11S1385	11	58	54%	1.19

LOD scores

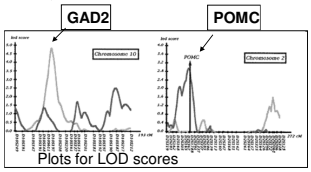
BMI results

Serum leptin levels

Affected sib pairs: sharing both parental alleles

Note peaks are over large intervals: candidate genes selected on functional information

Adapted from Hager et al., 1998, Nature genetics



Plots for LOD scores

Can we improve our data?

- Turn to quantitative aspects of discontinuous traits
  - These are the **endophenotypes**
  - E.g diabetes: BMI, serum leptin levels, serum insulin levels, triglyceride levels etc....
- Recognise that complex diseases show genetic heterogeneity, and work with multiple pedigrees
- Look at the bigger picture: can results be grouped into genetic pathways?

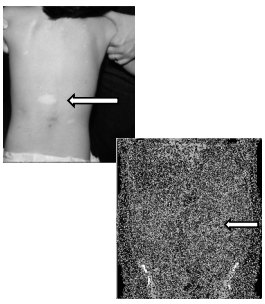
Genetic Case studies II

Taking a step back to some 'single' gene disorders.....

- Tumour with incomplete penetrance.
- This illustrates the problems associated with finding disease loci for conditions that do not show classical Mendelian behaviours.
  - Tuberous Sclerosis (TSC) – an example of (genetic) locus heterogeneity

Tuberous Sclerosis (Complex)

- Autosomal dominant (estimates 1/6,000 to 1/17,000 births
- 2/3 *de novo* mutation; 1/3 genetic
- Genetic cases: pedigrees show incomplete penetrance, with increasing severity over successive generations
  - (skin lesions in parents/grandparents, through to tumours and severe MR in probands)
- Multisystemic (brain, kidney, skin)
- Epilepsy, MR, autism
- Hamartomas (tumour like lesions)



For a review see JRW Yates European Journal of Human Genetics (2006) ***14,*** *1065-1073.*

Images from D.N Franz, Biologics. 2013; 7: 211–221.

Exploring the Genetics

- 1980s- start to look at multiple small pedigrees
- Contradictory results. More than one chromosome?
- Meta-analysis, no positive LOD score

- **Genetic Heterogeneity (**in this case **locus heterogeneity)** (more than one locus→ same disease phenotype)

- Linkage to 9q- *ABO blood group used to initially confirm this*
- Remove all families with evidence of linkage to 9q
- Reassess families not linked to 9q

Breakthrough

- Some families with TSC had other conditions linked to chr 16 (e.g. polycistic kidney disease)
- MS screen indicated LOH on chr 16
- Overlapping LOH intervals defined a strong candidate region
- Genomic clones from interval characterised
- Gene sequenced! Called tuberin.

- Similar approach for TSC1 on chr 9, but fewer families with LOH. Gene called hamartin.
- Both genes behave like tumour supressors

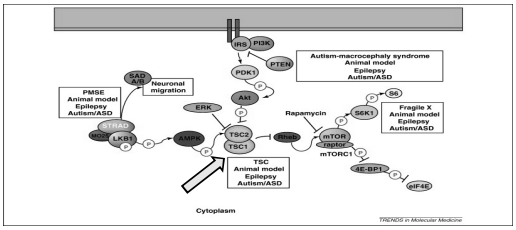


Figure 2 Schematic depicting multiple sites for mTORC1 dysregulation (phosphoactivating and phosphoinhibitory effects) that are linked with human neurological diseases ('mTORopathies'), including epilepsy and autism/autism spectrum disorder (ASD).

Peter B. Crino

mTOR: A pathogenic signaling pathway in developmental brain malformations

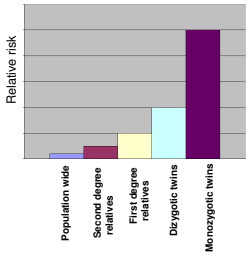
Trends in Molecular Medicine Volume 17, Issue 12 2011 734 - 742

http://dx.doi.org/10.1016/j.molmed.2011.07.008

Parent-of-origin effects

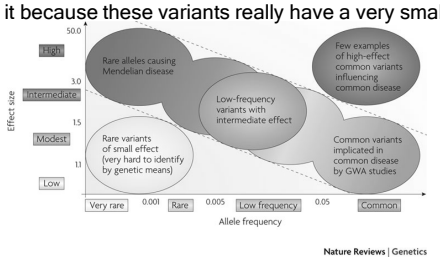
- Know this occurs for imprinted diseases
  - Again, gene deletion syndromes (=pathogenic CNVs) reveal which genes or regions of the genome are involved
- Don't know all imprinted loci
- Parent-sib trios on SNP arrays may reveal more about which loci have parent-of origin effects
  - Some already found in breast cancer, basal cell carcinoma, type 2 diabetes, autism (*Kong et al, 2009, 2012*)
- Can use standard SNP array techniques to decipher parental contributions where methylation is involved in the imprinting process

Complex phenotypes: has GWAS really delivered?



- Most complex disease can be shown to have a genetic component based on relative risk studies
- Most complex diseases of interest have high heritability estimates
- But most of the loci found account for very little of this heritability
- Why?

Is it because these variants really have a very small effect?



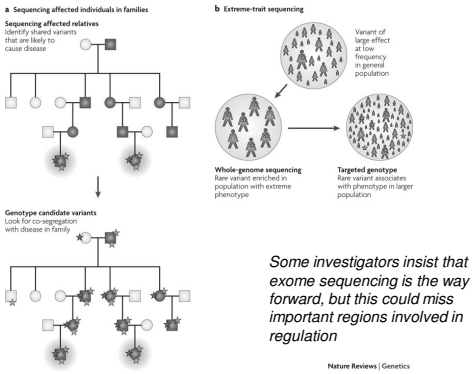
Nature Reviews | Genetics

Are other variants more important?

Variant class	Minor allele frequency	Implications for analysis
Very common	Between 5 and 50%	Amenable to association analysis using current genome-wide association methods
Less common	Between 1 and 5%	Amenable to association analysis using variants catalogued in the 1000 Genomes Project
Rare (but not private)	Less than 1% but still polymorphic in one or more major human populations	Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families
Private	Restricted to probands and immediate relatives	Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants

Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - Model? Rare variants and common diseases?



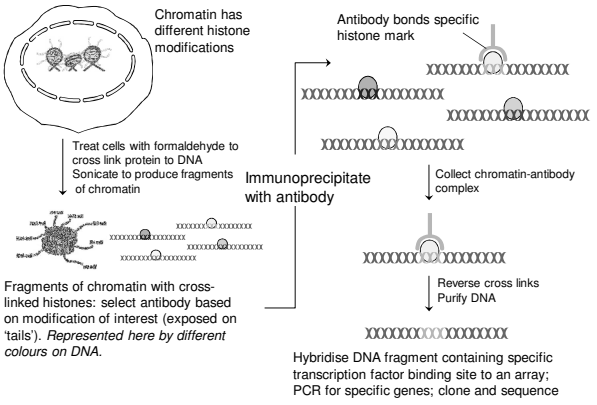
Some investigators insist that exome sequencing is the way forward, but this could miss important regions involved in regulation

Nature Reviews | Genetics

Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - Histone modifications

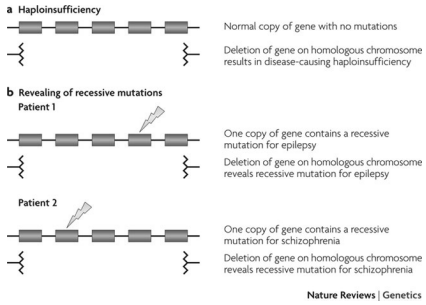
Chromatin **i**mmuno**p**recipitation: ChIP



Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - Copy number changes

What does a CNV reveal?

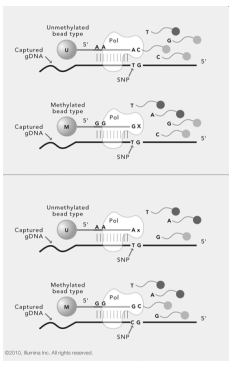


Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - DNA modifications (methylation/ hydroxymethylation: *these may be important for gene-environment interactions*)

DNA methylation studies

- Relatively easy to do
- Variation on the SNP array method
- Use bisulphite modification to chemically convert unprotected (unmethylated) C bases to uracil (read as T in sequencing stage)
- Score each by extending a single base **after** the one that is being investigated
  - If mismatch, no extension
- Ratio of the signals from methylated: unmethylated gives an estimate of DNA methylation at the site chosen
  - Concentrated on promoters and other runs of CpG dinucleotides

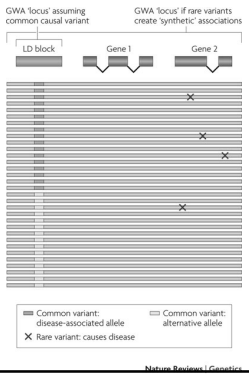


Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - Synthetic/ fortuitous association (*essentially a false positive*)

'Synthetic' or fortuitous association

- Here, an association picks up a block of linkage disequilibrium, but the actual mutations lie in another gene
- The association is **fortuitous**, since most of the mutations (all independent) are on chromosomes with the same haplotype at the LD block, and so share one of the alleles being scored more often than the other. (*The actual gene may lie in a different LD block, and therefore on multiple haplotypic backgrounds within that block - no association with adjacent SNP detected*)



Are we missing something else?

- This is being called the 'missing heritability' problem
- What other options should be considered?
  - Gene-gene interactions

Gene-Gene interactions: in the case of a single gene disorder

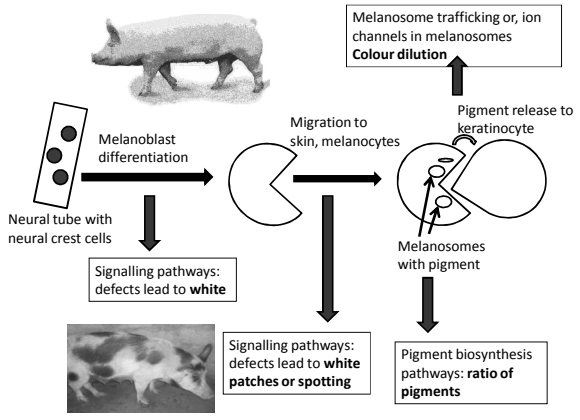
- Even Mendelian disorders are modified by environment and genetic background: this table shows gene alleles that are modifiers of sickle cell phenotypes. How much more complex might polygenic disorders be?

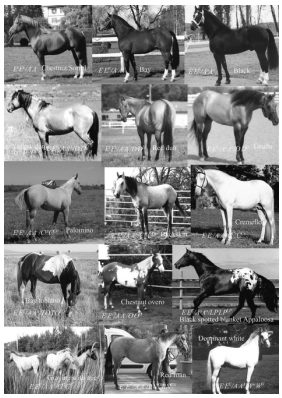
Subphenotype	Gene/SNP marker*	Effect	Reference
Stroke	*VCAM1*/G1238C	Protective†	Taylor *et al* (2002a)
	VCAM1/T-1594C	Permissive	Hoppe *et al* (2004)
	ILAR/S503P	Permissive	Hoppe *et al* (2004)
	TNFA/G-308S	Protective	Hoppe *et al* (2004)
	LDLR/Ncol +/−	Protective	Hoppe *et al* (2004)
	ADRB2/Q/27E	Protective	Hoppe *et al* (2004)
	AGT/AG repeats	Permissive	Tang *et al* (2001)
	HLA genes	Protective; permissive	see text
Osteonecrosis	*MTHFR*/C677T	Permissive, but questionable	De Castro *et al* (1998); Kutlar *et al* (1998); Zimmerman *et al* (1998); Zimmerman and Ware (1998)
Acute chest syndrome	*NOS3*/T-786C	Permissive	Sharan *et al* (2004)
	NOS1/AAT repeats	Permissive	Sullivan *et al* (2001)
Cholelithiasis	*UGT1A*/promoter repeats	7/7, Permissive	Passon *et al* (2001); Fertrin *et al* (2003)
Priapism	*KL*	Permissive	Nolan *et al* (2004)

From Steinberg, British Journal Haematology (2005); 129 p465

Classical epistasis: Coat colour

- Red pigment (pheomelanin), black pigment (eumelanin)
- MC1R polymorphism determines ratio of the pigments
- Polymorphic variants at other genetic loci conflict with genes governing melanin production: effects are **epistatic**, as they modify the phenotype predicted





A=Agouti (ASIP) recessive black
E= extension
(MC1R), recessive, epistatic over black

D= Dun epistatic over all colours

C=Cream epistatic to chestnut
Champagne, epistatic to all colours
Cream and champagne are both mutations in ion channels

Z=silver (PMEL17) epistatic to black

W=white (KIT)
O =overo (EDNRB)
LP= leopard spotting/ appaloosa (TRPM1, a calcium channel)
All of these are epistatic to basic coat colours

(*Other genetic variants are involved in white spotting. If you are interested see the paper*)

Journal of Animal Breeding and Genetics
Volume 126, Issue 6, pages 415-424, 12 NOV 2009 DOI: 10.1111/j.1439-0388.2009.00832.x
http://onlinelibrary.wiley.com/doi/10.1111/j.1439-0388.2009.00832.x/full#f1

So how should we use GWAS data?

- Association studies give us real clues, just as single gene disorders and contiguous gene deletion syndromes with overlapping phenotypes give us clues about polygenic diseases
- The estimated contributions of the alleles don't match expectations. But the locus could be correct, it's just that the mutations are not shared across individuals as a result of a common ancestral mutation (no common LD or haplotype blocks)
- Take all the cumulative evidence and use deep sequencing to find the missing DNA variants
- Just sequence in the first place!
- *Don't forget to look at epigenetic changes and parent-of-origin effects as additional layers of complexity*

Personalised medicine: using SNP data in treating the patient

- Pharmacogenetics and pharmacogenomics are key areas of research
- Need to know how drugs are metabolised
- Do genetic variants alter efficacy?
  - Toxicology
  - Side effects
  - Metabolites
- Cancer
  - Treatments target cancer cells based on presence of biomarkers (goes alongside functional analysis)
  - Need to know tumour biology
  - Need to know patient genetics for specific genes that may interfere with treatment

See chapter 19 in Strachan and Read for a fuller discussion of the above, plus the pros and cons of population testing for common diseases, such as breast cancer